SURVEY: A PLATFORM TO EXPLORE EMERGING VALUE PERSPECTIVES IN ALMS' BEHAVIORS

A PREPRINT

Steve J. Bickley, Ho Fai Chan, Bang Dao, Benno Torgler, Son Tran

Panalogy Lab
Brisbane, 4000
steven.bickley@panalogy-lab.com

July 10, 2023

ABSTRACT

This paper presents our work on Survey, a platform for analyzing augmented language models's (ALMs) emergent alignment behaviours through their attitude and value perspectives.

***K*eywords** Social Science • Social AI • Large Language Models • Augmented Language Models • Alignment Research • Survey Method • Experimental Method

1 Motivations

We built Survey withe the following motivations in mind:

- AI alignment in complex social context is **important**, especially when AI systems make decisions or assist people in making decisions in complex settings where there may be no true "right" answer (i.e., where "right" is contextual to the specific information of the real-world situation).
- Doing survey and experiments, despite shortcomings, is still a **widely** adopted method to study social behaviors, including alignments [1]. Language model evaluation frameworks are essentially survey/experiment models [3].
- Applying the survey approach to systematically analyzing ALMs' alignment behaviors is therefore desirable. ALMs are the most advanced AI systems with general reasoning capability, and the have been increasingly operating in complex social context with emergent behaviors not expected before.
- Being able to learn from ALM's feedback to **improve** survey/experiment design is a great application of ALMs, thus helping researchers to construct quality survey frameworks at a fraction of the resources and time that would be otherwise required when humans have to do everything.
- Survey is the **first** platform delivering robust results for the above value propositions.

2 Introduction

Our new platform, **Survey**, aims to explore emerging value perspectives in Augmented Language Models (ALMs) [4] through decision premises and contextuality in complex social settings. We chose to focus on complex social settings because these environments present a wide array of situations and dilemmas that help in assessing the multifaceted and nuanced value judgments that ALMs have to make in the real world. As a result, we can better understand their decision-making mechanisms and the underlying value systems they rely on.

Our approach to exploring ALM's behaviors, and their value perspectives, is based on presenting the models with complex sets of survey questions and experimental settings then analyzing their responses under in various contexts. We opted for a survey-based approach as it has been proven to be an effective method for ALM probing and value elicitation [6]. Value elicitation of humans has always been an important research avenue [8], as an individual's identity and values often manifest in the choices and decisions they make. Given the increasing role of ALMs as agents interacting with humans as principals (i.e., principal-agent relationship in nature), it becomes crucial to comprehend the underlying values of these models and their implications for recommendations and decision-making processes, which may or may not align with those of human principals. By allowing the models to respond to a series of contextualized questions, we can gain valuable insights into how they navigate ethical scenarios and prioritize differing values for different principal profiles (e.g., age, race, country of residence, etc).

The first key aspect we address is the question of why understanding ALM's value perspectives matters in alignment research. By examining the context-dependent values that drive ALMs' behaviours, we gain valuable insights into the ethical and moral frameworks and traces that shape their decision-making processes (as well as our own). These insights can help us align ALMs with (context-specific) human values and promote responsible AI development [9].

Defining measures of values is another crucial element of our platform, Survey. We recognize that values are multi-faceted and subjective, making their quantification challenging. Through rigorous research and analysis, we aim to develop and translate robust methodologies to define and measure values in the context of ALMs. This will provide a solid foundation for studying and understanding their behavior patterns in complex social settings.

To effectively explore emerging value perspectives in ALMs' behaviors, we employ state-of-the-art (SOTA) ALM simulation techniques. By simulating ALMs in realistic scenarios, we can observe their decision-making processes and identify any evolving biases or value systems. This knowledge is crucial for staying ahead of potential risks and ensuring the responsible development of AI technologies.

The potential applications of our platform extend beyond the fields of behavioural economics, cognitive psychology, or market research. We envision leveraging the insights gained to help drive social AI alignment and design in research, industry and community alike. By aligning AI systems with human values, we can (co-)create social AI systems that complement and enhance lives, benefiting society as a whole.

In conclusion, our platform Survey offers a unique opportunity to explore and understand the emerging value perspectives in ALMs' behaviors. By addressing the questions of why world values matter, defining measures of values, and simulating ALMs in complex social settings, we can unlock insights that will help shape the future of AI development. With a focus on social AI alignment and design, we aim to create long-lived, complex social AI agents and systems that align with human values, promoting a responsible and beneficial integration of AI into industry and society.

3 Platform Design

To be continued…

4 Applications

With the emergent capabilities of Large Language Models (LLMs) and the ongoing development of Augmented Language Models (ALMs), understanding the underlying premises that govern their behaviors and decision-making processes becomes increasingly crucial, especially when they operate in complex social settings and have real-world impact. This understanding is essential to ensure the alignment of social AI with human values.

ALMs, which are essentially built on top of pre-trained LLMs like gpt-4, incorporate various elements such as retrieval plug-ins, different learning techniques (few-shot), diverse prompting methods (such as chain-of-thought, self-model, and contextual prompts), functional coding, and integration with other modalities like voice, vision, and sound [10]. Additionally, future iterations of ALMs are expected to incorporate different AI techniques such as reinforcement learning and symbolic logic, enhancing their knowledge organisation, reasoning, and learning capabilities.

Researchers have been quick to recognize the potential of LLMs as valuable tools to study and probe the human mind and society[12], given their training on vast amounts of human data and their ability to generate human-like text. Others have discussed their potential in simulating human subjects [15]. Consequently, researchers from various disciplines such as behavioral economics, cognitive psychology, social psychology, linguistics, and others have proposed different tools, surveys, and methodologies to investigate LLMs' behavior and decision-making processes. However, the procedures and tuning of LLMs (e.g., temperature, context window, prompt context and structure) for judgment and evaluation of alignment are not yet standardized or consistently applied across studies. Moreover, digital literacy and programming skills continue to present significant obstacles for many researchers, particularly those in behavioral economics and the social sciences.

Considering the fast-paced nature of research and development in AI at the moment, it is essential to also extend our focus beyond pre-trained LLMs and consider the emergent capabilities and value systems of ALMs within various different social contexts. ALMs are increasingly augmented with additional tools and various prompting techniques, spanning different context windows and incorporating other modalities. Furthermore, these ALMs are now actively performing real-world actions. Calling a tool in the context of ALMs often involves having an impact on the virtual or physical world and observing the resulting effects, which are typically integrated into the ALM's ongoing context. Moreover, ALMs are increasingly engaging in delegate actions such as carrying out transactions on our behalf or responding to customer queries and emails in human-like ways.

By acknowledging the advancements in ALMs and the complex nature of their interactions with the world, we can gain a comprehensive understanding of the premises underlying their behaviors and decision-making processes. This knowledge is crucial for ensuring the development of responsible and aligned social AI systems that reflect human values for the benefit of all humankind.

To facilitate this exploration in an easy and intuitive manner, we have developed a social science research platform called Survey . It empowers researchers to investigate the behaviors and decision-making of LLMs and ALMs in a robust and systematic way, using an easy-to-use, click-and-play online interface. By simulating decision-making across a spectrum of randomised agent demographic attributes (e.g., age, gender, education level, personality, etc), Survey provides a unique platform to investigate even the most sensitive and taboo social science topics (e.g., end-of-life decisions, domestic violence, abortion, etc). By probing these areas in a simulated environment, we leverage the potential of ALMs to explore sensitive topics (e.g., health, social, economic, ethical, etc) in a safe and ethical environment.

5 Future Development

A key limitation of the social context of our simulation is that context is provided in a randomized manner (uniform distribution), and using a generic story template to build up the context of the agent (e.g., you are <AGE>, your personality is <BIG 5 PROFILE>, and reside in <LOCATION>). We do not draw on or attempt to simulate human subjects from demographic backgrounds of past survey respondents, as in e.g., [16]. Note, this approach also means that sometimes we end up with "interesting" agent profile combinations that may seldom present in the real world (e.g., a male lesbian).

As an initial proof-of-concept, we have naturally had to limit the number of context variables available in the Survey platform. A key next step is to allow the addition and customization of top-level profile attributes as well as their options (i.e., an extension to the existing capability for customising options within the existing set of top-level demographic attributes like age, gender, etc.). For example, to create new top-level attributes like employment status, rural vs urban residence, mental health status, occupational industry, marital/family status, etc.

Other future development in the pipeline for the Survey platform:

- Customisation of the "analysis" parameter/prompt to allow users to directly enter the prompt instructions to customise how the ALMs critically analyse the questions in "critic" mode.
- Parallelisation of OpenAI API queries to reduce simulation times.
- User interface - add more detailed progress indicator, add output visualisation

References

[1] A. Bhattacherjee, "Social Science Research: Principles, Methods, and Practices."

[2] "Openai/evals: Evals is a framework for evaluating LLMs and LLM systems, and an open-source registry of benchmarks." Available: https://github.com/openai/evals

[3] A. Srivastava *et al.*, "Beyond the Imitation Game: Quantifying and extrapolating the capabilities of language models." arXiv, Jun. 2022. doi: 10.48550/arXiv.2206.04615.

[4] G. Mialon *et al.*, "Augmented Language Models: A Survey." arXiv, Feb. 2023. doi: 10.48550/arXiv.2302.07842.

[5] A. Arora, L.-A. Kaffee, and I. Augenstein, "Probing Pre-Trained Language Models for Cross-Cultural Differences in Values."

[6] M. Binz and E. Schulz, "Using cognitive psychology to understand GPT-3," *Proceedings of the National Academy of Sciences*, vol. 120, no. 6, p. e2218523120, Feb. 2023, doi: 10.1073/pnas.2218523120.

[7] C. Haerpfer, R. Inglehart, A. Moreno, C. Welzel, K. Kizilova, and J. Diez-Medrano, "World values survey: Round seven–country-pooled datafile."

[8] G. Hofstede, "Culture's recent consequences." Product & Systems Internationalisation, Inc., 2005. Available: 3-4

[9] A. Tamkin, M. Brundage, J. Clark, and D. Ganguli, "Understanding the capabilities, limitations, and societal impact of large language models," doi: 10.48550/arXiv.2102.02503.

[10] G. Mialon *et al.*, "Augmented Language Models: a Survey."

[11] A. Korinek, "Language Models and Cognitive Automation for Economic Research," 2023.

[12] M. Miotto, N. Rossberg, and B. Kleinberg, "Who is GPT-3? An Exploration of Personality, Values and Demographics."

[13] G. Aher, R. I. Arriaga, and A. T. Kalai, "Using Large Language Models to Simulate Multiple Humans and Replicate Human Subject Studies."

[14] J. J. Horton, "Large Language Models as Simulated Economic Agents: What Can We Learn from Homo Silicus?"

[15] J. S. Park, L. Popowski, C. J. Cai, M. R. Morris, P. Liang, and M. S. Bernstein, "Social Simulacra: Creating Populated Prototypes for Social Computing Systems."

[16] L. P. Argyle *et al.*, "AI Chat Assistants can Improve Conversations about Divisive Topics."